Estimating corporate income tax gaps: A bottom-up approach for Slovakia

M. Chudý¹  R. Gábik²  J. Bukovina¹

¹Institute for financial policy, Ministry of Finance SR
²Financial Administration SR

OECD, June 27th

This is an ongoing work and the list of coauthors is not yet complete. The opinions are those of the authors and do not necessarily reflect the official policy or position of any other agency or organization.
Tax Gap is the difference between the amount of taxes collected and the amount that should have been collected if all taxpayers had followed all rules and regulations to the letter.
Agenda

- Introduction
  - Tax enforcement
  - Bottom-up estimator
- Methods
- Data
  - individual characteristics of all firms,
  - operational audit data for a sample of firms,
  - descriptive statistics,
  - recommendations for the financial administration.
- CIT Gap estimates
  - implementation
  - bottom-up vs top-down
- Outlook
Tax enforcement:
- $\uparrow$ tax compliance = $\uparrow$ revenues

GOALS:
- $\downarrow$ legislation uncertainty
- $\downarrow$ compliance burden
- $\downarrow$ legislative loopholes
- $\uparrow$ audits
- $\downarrow$ non-deliberate errors,
- $\downarrow$ ghost firms,
- $\downarrow$ deliberate non-compliance,
- $\downarrow$ evasion.

MEANS:
- $\uparrow$ number of audits
- $\uparrow$ targeting of audits
- $\uparrow$ risk for all evaders
- $\uparrow$ risk for some\(^1\) evaders
- $\uparrow$ costs & revenues
- $\uparrow$ revenues.

BY-PRODUCT:
- data to assess the entire population *for free.*
- tailor-made inputs for design of tax policies.

\(^1\)Targeted audits select firms based on propensity and scale of non-compliance therefore some small firms face no risk.
Bottom-up approach
-For fixed tax period (year):

- Population: $N$ firms $\rightarrow$ set of $K$ characteristics $X = (X_1, \ldots, X_K)$.
- Subset: $n \ll N$ $\rightarrow$ we observe levels of non-compliance $Y_1, \ldots, Y_n$ determined by audits.
- We predict missing $Y$’s using $E(Y|X)$ from a linear model.
  - Classical approach $\rightarrow$ the $n$ audits are random and representative in terms of characteristics $X$.
  - In reality $\rightarrow$ $\mathcal{L}(Y|X) \neq \mathcal{L}(Y|X, \text{selected for audit})$.

- Standard tools for censored data and selection bias:
  - Tobit (Tobin, 1958; Hanlon et al., 2007),
  - Heckman (Heckman, 1979; Erard and Feinstein, 2007),
  - Matching (Rosenbaum and Rubin, 1983; Nicolay, 2013).
Methods
- Heckman (1979): “Sample selection bias ≈ problem of omitted predictor.”

Assume that \((Y_S^*, Y_O^*) \in \mathbb{R}^2\) satisfy

\[
\begin{align*}
Y_S^* &= \beta_S^T X_S + \varepsilon_S, \quad (1) \\
Y_O^* &= \beta_O^T X_O + \varepsilon_O, \quad \text{where} \\
\begin{bmatrix} \varepsilon_S \\ \varepsilon_O \end{bmatrix} &\sim \mathcal{N}_2 \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \sigma \\ \rho \sigma & \sigma^2 \end{bmatrix} \right). \quad (3)
\end{align*}
\]

For \(j = 1, \ldots, N\), we observe realizations of \((Y_{Sj}, Y_{Oj}, X_{Sj}^T, X_{Oj}^T)\), where

\[
\begin{align*}
Y_S &= \begin{cases} 0 & Y_{Sj}^* \leq 0, \\ 1 & \text{else}, \end{cases} \quad (4) \\
Y_O &= \begin{cases} \text{NA} & Y_S = 0, \\ Y_{Oj}^* & Y_S = 1. \end{cases} \quad (5)
\end{align*}
\]
Methods
- Heckman’s sample bias correction

- A firm is selected for audit on event $Y_s^* > 0 | X_s = x_s$.
- We observe $Y_s^*$ indirectly through binary $Y_s^2$.
- Under the Probit model, $P(Y_s = 1 | X_s) = \Phi(\beta X_s)$.
- Heckman (1979) proposes a two-step OLS later MLE.
- Development $\rightarrow$ non-parametric approach (Powell, 1984).
- Heckman in the R (Toomet and Henningsen, 2008)\(^3\).
- MLE requires good starting values (Chen and Zhou, 2010).

\(^2\)If $Y_O = Y_s$ $\rightarrow$ use Tobit model (Hanlon et al., 2007).

\(^3\)The package implements both OLS and MLE with multiple optimisation methods. Beside numeric output, it also implement interval regression.
Methods
- Examples of empirical studies and extensions:

- Erard and Ho (2001) → optimize firm’s compliance facing audit risk.
- Beer et al. (2019); EC (2018) → cross-country meta-studies and technical reports
- Hanlon et al. (2007) → Tobit model

---

4 In every case-study, the details and specifics of the individual country require special attention.
Methods
- Going beyond Heckman’s model

- We observe 0 non-compliance in 50% of audits.
- Our intuition → auditors might overlook some small deficiencies (below certain threshold $T$)
- We propose a model for twice-censored data

So in addition to (1) and (2) we have

$$Y_S = \begin{cases} 
0 & Y^*_S \leq 0, \\
1 & \text{else,}
\end{cases}$$

$$Y_O = \begin{cases} 
\text{NA} & Y_S = 0, \\
0 & Y_S = 1, \text{ and } Y^*_O \leq T, \\
Y^*_O & Y_S = 1, \text{ and } Y^*_O > T,
\end{cases}$$

and where $0 \leq T < \infty$ is the known threshold. We can estimate the unknown parameters $\theta = (\beta_S^T, \beta_O^T, \rho, \sigma)$ using the ML.
Data
- Individual firm characteristics for the entire population

The complexity of the firm-level data is threefold:

- **Serial** → the period 2014-2016.
- **Cross-sectional** → entire population of active firms in Slovakia for each year\(^5\).
- **Individual** → non-financial (firm’s legal, sectoral, geographical and social profile) and financial (mostly fields from the financial statement as well as CIT and VAT returns and respective tax audits) characteristics.

---

\(^5\)FD contains only legal persons that are liable to CIT. Some entities such as self-employed liable to personal income tax (PIT) are not present.
Data
- Individual non-compliance for the audited set

- key input for the bottom-up CIT gap estimation,
- the results of CIT audits are available with a delay of nine months on average but sometimes years,
- currently, finalized audits up to the tax year 2016,
- the amount of usable audit data is small compared to the population size → merge audit data across the entire range of tax periods (2014-2016),
Data
-Cleaning procedure

- raw panel → 900 000 unique firm-year entries,
- merging and simplification of similar variables from different sources (size, NACE sector),
- from 117 we keep only 9 most populated: Number of employees, NACE Sector, Ownership, Administrative location, Personal costs, Profit/Loss, Net assets, Value added.
- for the sake of comparison with top-down → exclude financial and non-profit organisations.
- exclusion of certain types of CIT audits: “by tools”, “tax licence”, “selected large firms”, “tax loss reduction”.

## Data
- Cleaning procedure

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) raw data</td>
<td>308 363</td>
<td>309 111</td>
<td>307 768</td>
<td>4 776</td>
</tr>
<tr>
<td>(i) omit missing tax ID</td>
<td>236 481</td>
<td>237 521</td>
<td>233 446</td>
<td>4 776</td>
</tr>
<tr>
<td>(ii) omit missing values</td>
<td>186 970</td>
<td>187 868</td>
<td>198 005</td>
<td>4 776</td>
</tr>
<tr>
<td>(iii) omit non-profit, banks, etc.</td>
<td>185 525</td>
<td>186 545</td>
<td>197 207</td>
<td>4 761</td>
</tr>
<tr>
<td>(iv) filter only audits conducted for tax-period 2014-2016</td>
<td></td>
<td></td>
<td></td>
<td>3 263</td>
</tr>
<tr>
<td>(v) screen for duplicitous audits over successive years</td>
<td></td>
<td></td>
<td></td>
<td>2 464</td>
</tr>
<tr>
<td>(vi) augment audits with selected 9 individual characteristics</td>
<td></td>
<td></td>
<td></td>
<td>1 487</td>
</tr>
<tr>
<td>(vii) omit residual audits without full assessment of accounting</td>
<td></td>
<td></td>
<td></td>
<td>1 344</td>
</tr>
<tr>
<td>(viii) omit residual firms subjected to special tax legislation</td>
<td></td>
<td></td>
<td></td>
<td>1 329</td>
</tr>
<tr>
<td>(ix) omit audits targeted at “tax licence”</td>
<td></td>
<td></td>
<td></td>
<td>1 294</td>
</tr>
<tr>
<td>(x) omit audits resulting in negative tax deficiencies</td>
<td></td>
<td></td>
<td></td>
<td>1 292</td>
</tr>
<tr>
<td>(xi) omit the audits resulting in reduction of tax loss</td>
<td></td>
<td></td>
<td></td>
<td>1 126</td>
</tr>
</tbody>
</table>
### Size

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>164 240</td>
<td>164 285</td>
<td>174 226</td>
<td>225</td>
<td>349</td>
<td>67</td>
<td>641</td>
</tr>
<tr>
<td>Small</td>
<td>17 359</td>
<td>18 223</td>
<td>18 899</td>
<td>140</td>
<td>174</td>
<td>50</td>
<td>364</td>
</tr>
<tr>
<td>Medium</td>
<td>3 247</td>
<td>3 360</td>
<td>3 383</td>
<td>43</td>
<td>39</td>
<td>25</td>
<td>107</td>
</tr>
<tr>
<td>Large</td>
<td>679</td>
<td>677</td>
<td>699</td>
<td>8</td>
<td>6</td>
<td>-</td>
<td>14</td>
</tr>
</tbody>
</table>

### Region

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bratislava</td>
<td>61 552</td>
<td>62 397</td>
<td>66 150</td>
<td>78</td>
<td>143</td>
<td>29</td>
<td>250</td>
</tr>
<tr>
<td>Trnava</td>
<td>16 163</td>
<td>16 284</td>
<td>17 355</td>
<td>44</td>
<td>67</td>
<td>17</td>
<td>128</td>
</tr>
<tr>
<td>Trencin</td>
<td>14 813</td>
<td>14 740</td>
<td>15 622</td>
<td>53</td>
<td>56</td>
<td>23</td>
<td>132</td>
</tr>
<tr>
<td>Nitra</td>
<td>20 279</td>
<td>20 264</td>
<td>21 162</td>
<td>37</td>
<td>59</td>
<td>11</td>
<td>107</td>
</tr>
<tr>
<td>Zilina</td>
<td>19 151</td>
<td>19 193</td>
<td>20 345</td>
<td>56</td>
<td>88</td>
<td>17</td>
<td>161</td>
</tr>
<tr>
<td>B. Bystrica</td>
<td>16 629</td>
<td>16 617</td>
<td>17 625</td>
<td>66</td>
<td>61</td>
<td>11</td>
<td>138</td>
</tr>
<tr>
<td>Presov</td>
<td>16 901</td>
<td>16 977</td>
<td>17 920</td>
<td>49</td>
<td>56</td>
<td>17</td>
<td>122</td>
</tr>
<tr>
<td>Kosice</td>
<td>20 030</td>
<td>20 065</td>
<td>21 028</td>
<td>33</td>
<td>38</td>
<td>17</td>
<td>88</td>
</tr>
</tbody>
</table>

### Sector

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accomodation</td>
<td>7 078</td>
<td>7 125</td>
<td>7 620</td>
<td>16</td>
<td>26</td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td>Agriculture</td>
<td>4 520</td>
<td>4 605</td>
<td>4 838</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Construction</td>
<td>19 464</td>
<td>19 559</td>
<td>20 897</td>
<td>64</td>
<td>120</td>
<td>53</td>
<td>237</td>
</tr>
<tr>
<td>Finance</td>
<td>1 974</td>
<td>2 072</td>
<td>2 389</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Industry</td>
<td>16 239</td>
<td>16 266</td>
<td>17 030</td>
<td>62</td>
<td>58</td>
<td>16</td>
<td>136</td>
</tr>
<tr>
<td>Information</td>
<td>10 828</td>
<td>11 250</td>
<td>12 056</td>
<td>11</td>
<td>16</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Others</td>
<td>14 957</td>
<td>15 377</td>
<td>16 371</td>
<td>14</td>
<td>11</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>RealEstate</td>
<td>13 400</td>
<td>13 436</td>
<td>14 118</td>
<td>17</td>
<td>21</td>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>Specialized</td>
<td>29 639</td>
<td>30 134</td>
<td>32 082</td>
<td>30</td>
<td>42</td>
<td>6</td>
<td>78</td>
</tr>
<tr>
<td>Supporting</td>
<td>11 587</td>
<td>11 726</td>
<td>12 833</td>
<td>24</td>
<td>26</td>
<td>6</td>
<td>56</td>
</tr>
<tr>
<td>Transport</td>
<td>7 720</td>
<td>7 760</td>
<td>7 881</td>
<td>29</td>
<td>26</td>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>WholesaleAndRetail</td>
<td>48 119</td>
<td>47 235</td>
<td>49 092</td>
<td>141</td>
<td>216</td>
<td>31</td>
<td>388</td>
</tr>
</tbody>
</table>

### Ownership

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>domestic</td>
<td>156 946</td>
<td>158 623</td>
<td>168 864</td>
<td>337</td>
<td>468</td>
<td>120</td>
<td>925</td>
</tr>
<tr>
<td>foreign</td>
<td>28 579</td>
<td>27 922</td>
<td>28 343</td>
<td>79</td>
<td>100</td>
<td>22</td>
<td>201</td>
</tr>
</tbody>
</table>
## Descript.
### -summary II

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Audits</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>mean</td>
<td>median</td>
<td>max</td>
<td>min</td>
<td>mean</td>
<td>median</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>Net assets</td>
<td>-4 607 215</td>
<td>1 061 566</td>
<td>41 070</td>
<td>8 880 572</td>
<td>-67 972</td>
<td>2 026 693</td>
<td>561 848</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value added</td>
<td>-87 721 000</td>
<td>196 871</td>
<td>7 069</td>
<td>1 028 123</td>
<td>-74 078 946</td>
<td>376 602</td>
<td>79 573</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit and Loss</td>
<td>-79 495 440</td>
<td>54 618</td>
<td>1 808</td>
<td>666 000</td>
<td>-79 495 440</td>
<td>90 250</td>
<td>11 778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total revenues</td>
<td>-2 583 780</td>
<td>1 011 168</td>
<td>39 787</td>
<td>7 227 454</td>
<td>0</td>
<td>3 304 207</td>
<td>1 179 738</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal costs</td>
<td>-105 508</td>
<td>99 797</td>
<td>1 357</td>
<td>362 851</td>
<td>-86</td>
<td>251 460</td>
<td>35 752</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Relationships:
- ownership-profit
- sector-assets
- region-revenues
<table>
<thead>
<tr>
<th>Size</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>0</td>
<td>51 002</td>
<td>608</td>
<td>3 308 882</td>
<td>32 692 397</td>
</tr>
<tr>
<td>Small</td>
<td>0</td>
<td>40 487</td>
<td>630</td>
<td>3 758 730</td>
<td>14 737 325</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
<td>54 013</td>
<td>411</td>
<td>2 003 729</td>
<td>5 779 346</td>
</tr>
<tr>
<td>Large</td>
<td>0</td>
<td>146 628</td>
<td>290</td>
<td>1 095 228</td>
<td>2 052 795</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bratislava</td>
<td>0</td>
<td>30 825</td>
<td>176</td>
<td>2 003 729</td>
<td>7 706 333</td>
</tr>
<tr>
<td>Trnava</td>
<td>0</td>
<td>100 499</td>
<td>0</td>
<td>3 758 730</td>
<td>12 863 819</td>
</tr>
<tr>
<td>Trenčín</td>
<td>0</td>
<td>22 979</td>
<td>1 803</td>
<td>637 743</td>
<td>3 033 276</td>
</tr>
<tr>
<td>Nitra</td>
<td>0</td>
<td>104 823</td>
<td>3 353</td>
<td>3 251 660</td>
<td>11 216 114</td>
</tr>
<tr>
<td>Žilina</td>
<td>0</td>
<td>72 442</td>
<td>2 711</td>
<td>3 308 882</td>
<td>11 663 127</td>
</tr>
<tr>
<td>B. Banská Bystrica</td>
<td>0</td>
<td>32 301</td>
<td>695</td>
<td>863 642</td>
<td>4 457 598</td>
</tr>
<tr>
<td>Prešov</td>
<td>0</td>
<td>16 415</td>
<td>1 293</td>
<td>329 197</td>
<td>2 002 613</td>
</tr>
<tr>
<td>Košice</td>
<td>0</td>
<td>26 352</td>
<td>0</td>
<td>1 266 211</td>
<td>2 318 984</td>
</tr>
<tr>
<td>Sector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accommodation</td>
<td>0</td>
<td>10 828</td>
<td>319</td>
<td>146 143</td>
<td>487 278</td>
</tr>
<tr>
<td>Agriculture</td>
<td>0</td>
<td>54 211</td>
<td>6 307</td>
<td>529 694</td>
<td>867 372</td>
</tr>
<tr>
<td>Construction</td>
<td>0</td>
<td>40 980</td>
<td>2 387</td>
<td>3 251 660</td>
<td>9 712 154</td>
</tr>
<tr>
<td>Finance</td>
<td>13 397</td>
<td>13 397</td>
<td>13 397</td>
<td>13 397</td>
<td>13 397</td>
</tr>
<tr>
<td>Industry</td>
<td>0</td>
<td>35 250</td>
<td>542</td>
<td>1 201 415</td>
<td>4 794 035</td>
</tr>
<tr>
<td>Information</td>
<td>0</td>
<td>91 569</td>
<td>337</td>
<td>1 741 763</td>
<td>2 930 215</td>
</tr>
<tr>
<td>Others</td>
<td>0</td>
<td>8 286</td>
<td>64</td>
<td>140 634</td>
<td>232 020</td>
</tr>
<tr>
<td>Real Estate</td>
<td>0</td>
<td>36 138</td>
<td>0</td>
<td>1 468 864</td>
<td>1 626 231</td>
</tr>
<tr>
<td>Specialized</td>
<td>0</td>
<td>42 734</td>
<td>528</td>
<td>1 266 211</td>
<td>3 333 251</td>
</tr>
<tr>
<td>Supporting</td>
<td>0</td>
<td>72 212</td>
<td>2 608</td>
<td>1 095 228</td>
<td>4 043 864</td>
</tr>
<tr>
<td>Transport</td>
<td>0</td>
<td>16 928</td>
<td>665</td>
<td>153 752</td>
<td>1 083 386</td>
</tr>
<tr>
<td>Wholesale And Retail</td>
<td>0</td>
<td>67 368</td>
<td>132</td>
<td>3 758 730</td>
<td>26 138 663</td>
</tr>
<tr>
<td>Ownership</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>domestic</td>
<td>0</td>
<td>44 062</td>
<td>845</td>
<td>3 758 730</td>
<td>40 757 324</td>
</tr>
<tr>
<td>foreign</td>
<td>0</td>
<td>72 162</td>
<td>0</td>
<td>2 516 809</td>
<td>14 504 539</td>
</tr>
</tbody>
</table>
Data

-Recommendations for the financial administration

- Centralize the selection of entities (organizational change required).
- Increase the number of audits and minimize the number of “by-tools” controls.
- Introduce a standardized audit output that will include the reasons for selecting the entity; precisely defined indicators (requires internal processes change).
- Evaluate audits by the amount of tax paid after review (requires a change of the key performance indicator).
- Clean up the taxpayer’s register of active entities in cooperation with the Ministry of Justice (requires legislative changes, internal process guidelines).
CIT Gap estimates

- Implementation

- Clean and merge the audit data.
- Augment audits with selected individual characteristics.
- Transform variables to logarithms and screen for outliers.
- Draw 5000 non-audited firms and obtain balanced sample of 6126.
- Estimate a probit model using optimal weights (Manski and Lerman, 1977).
- Choose the best subset of predictors by LASSO (Tibshirani and Knight, 1999).
- Estimate the censored Heckman using log-likelihood.
- Predict the deficiency for the entire population of firms active in a particular tax period.
Estimating CIT Gap:
- Bottom-up vs top-down

![Graph showing CIT Gap over years](image-url)
## CIT Gap estimates

### Summary statistics

<table>
<thead>
<tr>
<th>Gap</th>
<th>total (% revenues)</th>
<th>median (Eur)</th>
<th>max (mil. Eur)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLS</td>
<td>9</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>Heckman</td>
<td>10</td>
<td>8</td>
<td>34</td>
</tr>
<tr>
<td>Heckman-censored</td>
<td>17</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td><strong>Statistical matching</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 factor strata</td>
<td>21</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Propensity score</td>
<td>282</td>
<td>254</td>
<td>296</td>
</tr>
<tr>
<td><strong>Top-down</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIT</td>
<td>29</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>VAT</td>
<td>32</td>
<td>32</td>
<td>28</td>
</tr>
</tbody>
</table>
### Descriptive Summary IV

#### Size

<table>
<thead>
<tr>
<th>Size</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
<th>total</th>
<th>% of Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro</td>
<td>3</td>
<td>3 006</td>
<td>567</td>
<td>3 199 594</td>
<td>493 792 835</td>
<td>98.77</td>
</tr>
<tr>
<td>Small</td>
<td>3</td>
<td>316</td>
<td>81</td>
<td>704 649</td>
<td>5 750 582</td>
<td>1.15</td>
</tr>
<tr>
<td>Medium</td>
<td>4</td>
<td>101</td>
<td>36</td>
<td>46 717</td>
<td>340 557</td>
<td>0.07</td>
</tr>
<tr>
<td>Large</td>
<td>-</td>
<td>63</td>
<td>16</td>
<td>17 041</td>
<td>42 491</td>
<td>0.01</td>
</tr>
</tbody>
</table>

#### Region

<table>
<thead>
<tr>
<th>Region</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
<th>total</th>
<th>% of Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bratislava</td>
<td>-</td>
<td>1 946</td>
<td>385</td>
<td>3 199 594</td>
<td>121 406 025</td>
<td>24.28</td>
</tr>
<tr>
<td>Trnava</td>
<td>6</td>
<td>3 739</td>
<td>576</td>
<td>2 092 554</td>
<td>60 886 205</td>
<td>12.18</td>
</tr>
<tr>
<td>Trenčín</td>
<td>4</td>
<td>2 630</td>
<td>372</td>
<td>1 149 240</td>
<td>38 762 957</td>
<td>7.75</td>
</tr>
<tr>
<td>Nitra</td>
<td>7</td>
<td>3 980</td>
<td>643</td>
<td>1 837 905</td>
<td>80 651 860</td>
<td>16.13</td>
</tr>
<tr>
<td>B. Bystrica</td>
<td>6</td>
<td>3 817</td>
<td>630</td>
<td>1 913 093</td>
<td>63 423 071</td>
<td>12.69</td>
</tr>
<tr>
<td>Presov</td>
<td>4</td>
<td>1 823</td>
<td>275</td>
<td>515 129</td>
<td>30 946 942</td>
<td>6.19</td>
</tr>
<tr>
<td>Košice</td>
<td>1</td>
<td>2 415</td>
<td>346</td>
<td>3 180 735</td>
<td>48 449 746</td>
<td>9.69</td>
</tr>
</tbody>
</table>

#### Sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
<th>total</th>
<th>% of Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodation</td>
<td>7</td>
<td>1 570</td>
<td>289</td>
<td>84 793</td>
<td>11 186 228</td>
<td>2.24</td>
</tr>
<tr>
<td>Agriculture</td>
<td>17</td>
<td>7 048</td>
<td>1 345</td>
<td>1 913 093</td>
<td>32 457 872</td>
<td>6.49</td>
</tr>
<tr>
<td>Construction</td>
<td>3</td>
<td>2 429</td>
<td>419</td>
<td>1 837 905</td>
<td>47 500 231</td>
<td>9.50</td>
</tr>
<tr>
<td>Finance</td>
<td>11</td>
<td>3 173</td>
<td>789</td>
<td>546 356</td>
<td>6 575 020</td>
<td>1.32</td>
</tr>
<tr>
<td>Industry</td>
<td>-</td>
<td>1 819</td>
<td>173</td>
<td>1 059 076</td>
<td>29 588 808</td>
<td>5.92</td>
</tr>
<tr>
<td>Information</td>
<td>2</td>
<td>2 747</td>
<td>578</td>
<td>877 231</td>
<td>30 907 890</td>
<td>6.18</td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
<td>2 012</td>
<td>229</td>
<td>1 616 641</td>
<td>30 937 737</td>
<td>6.19</td>
</tr>
<tr>
<td>RealEstate</td>
<td>3</td>
<td>1 321</td>
<td>306</td>
<td>299 529</td>
<td>17 753 598</td>
<td>3.55</td>
</tr>
<tr>
<td>Specialized</td>
<td>3</td>
<td>2 432</td>
<td>487</td>
<td>3 199 594</td>
<td>73 282 105</td>
<td>14.66</td>
</tr>
<tr>
<td>Supporting</td>
<td>10</td>
<td>3 876</td>
<td>859</td>
<td>976 613</td>
<td>45 449 577</td>
<td>9.09</td>
</tr>
<tr>
<td>Transport</td>
<td>1</td>
<td>1 862</td>
<td>238</td>
<td>509 381</td>
<td>14 445 347</td>
<td>2.89</td>
</tr>
<tr>
<td>WholesaleAndRetail</td>
<td>2</td>
<td>3 384</td>
<td>602</td>
<td>3 180 735</td>
<td>159 842 053</td>
<td>31.97</td>
</tr>
</tbody>
</table>

#### Ownership

<table>
<thead>
<tr>
<th>Ownership</th>
<th>min</th>
<th>mean</th>
<th>median</th>
<th>max</th>
<th>total</th>
<th>% of Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>domestic</td>
<td>1</td>
<td>2 477</td>
<td>412</td>
<td>2 148 126</td>
<td>392 886 394</td>
<td>78.59</td>
</tr>
<tr>
<td>foreign</td>
<td>-</td>
<td>3 834</td>
<td>595</td>
<td>3 199 594</td>
<td>107 040 071</td>
<td>21.41</td>
</tr>
</tbody>
</table>
Outlook

This year:
- Look at the individual creative accounting behaviour after 2014 (minimal tax has been introduced).
- Incorporate information about common ownership.

Next year
- Bottom-up CIT gap for large firms and multinationals.
- Bottom-up VAT gap.

Maybe once:
- Bottom-up CIT and VAT gap from the perspective of a firm optimizing tax compliance (Erard and Ho, 2001).
Au revoir et

Merci de votre attention!


References III


References IV


References V


CIT revenues: Effective tax rate
-Naive measure of performance for tax collection

2012-2017
- nominal GDP = 85 bil. Eur (average yearly growth 3.1%).
- state revenues = 33 bil. Eur (average yearly growth 4.6%)
- CIT revenues = 2.6 bil. Eur (average yearly growth 11%)
CIT revenues: dynamics by components

-based on Vyskrabka and Antalicova (2018)
Top-down estimator in a nutshell
-based on Ueda (2018) and Gabik and Motkova (2019)

- INPUT: National accounts data (+ individual tax returns)
- METHODS:
  - CIT: 1. GDP $\rightarrow$ GOS 2. GOS + differences in national vs financial accounting 3. + valid CIT legislation = potential CIT base.
- OUTPUT: Tax gap for entire population of firms (trend).
- CONCLUSION:
  - Reliability: for VAT (√), for CIT (???)
  - Pros: large scope and clarity.
  - Cons: no insight, no linkage with firm’s characteristics.
Idea behind bottom-up approach

- Levels of non-compliance in the population:
Idea behind bottom-up approach
- Level of non-compliance in a random sample
Idea behind bottom-up approach

- Level of non-compliance observed in audits

Level of noncompliance

- small
- medium
- large
- mixture
Idea behind bottom-up approach

- Application of standard linear model to biased sample
Best prediction

- We believe that $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ are associated.
- Predict a new copy $Y^*$ using $m(X^*)$ while minimizing
  \[
  E (Y^* - m(X^*))^2.
  \]  
  (8)
- The best prediction is the projection on the closed subspace spanned by $X^*$, i.e. $m_0 = E (Y|X^*)$, which is (mostly) unknown, since the joint distribution is unknown.
- Assume without loss of generality $X_1 = 1$.
- We typically use a linear proxy $m_{\text{lin}} = X^T \beta$, $\beta \in \mathbb{R}^p$.
- The best linear prediction is given by
  \[
  m_{\text{lin},0} = (E (y1), \ldots, E (yX_p)) \left( E (XX^T) \right)^{-1} X^*.
  \]
- Given data $Y, X$, where $(Y_i, X_i^T)^{T \ i.i.d.} \sim (Y, X^T)^T$, $i = 1, \ldots, n$, we use LS to get estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Note: $m(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}, p < n$, is measurable with finite 2$^{\text{nd}}$ moment and the matrix $X$ has full column rank $p$. 

36/23
Best prediction and causal inference based on model

- Correctly specified linear model

Typically we,

■ choose the best prediction among $2^p$ candidates,
■ infer the causal effect of each $X_i, i = 1, \ldots, p$ on $E(Y|X)$.

To have a systematic approach, we impose further structure.
Assume that random sample data $(Y, X)$ satisfy a linear model

$$E(Y|X) = X\beta \quad \text{and} \quad \text{var}(Y|X) = \sigma^2 I_n.$$ 

Then for the error terms $\varepsilon = Y - X\beta$

■ $E(\varepsilon|X) = 0_n$ and $\text{var}(\varepsilon|X) = \sigma^2 I_n$
■ Gauss-Markov: $\hat{m}_{LS} := X\hat{\beta}$ is BLUE of $E(Y|X)$.
■ (8) reaches the optimal value $\sigma^2 + \text{var}(\hat{m}_{LS}) = \sigma^2(p + 1)$.
■ $\hat{\beta}$ for $n \to \infty$ reaches Cramer-Rao lower bound.
Best prediction and causal inference based on model

-Misspecified linear model: Omitted important regressors

Assume that we have data \((Y, X = (U, V))\), with \(\text{rank}(U) + \text{rank}(V) = p\) and two nested candidate linear models

(I.) \(Y \mid X \sim (U\beta, \sigma^2 I_n)\),

(II.) \(Y \mid X \sim (U\beta + V\gamma, \sigma^2 I_n)\).

Irrespective of which model is correct:

- \(\text{var} (\hat{m}_{LS}(II.)) - \text{var} (\hat{m}_{LS}(I.)) \geq 0_n\),
- \(\text{var} (\hat{\beta}(II.)) - \text{var} (\hat{\beta}(I.)) \geq 0_n\).

If (II.) is correct, but, for any reason, we estimate (I.) instead:

- \(E (\hat{\beta} - \beta \mid X) = (U^TU)^{-1} U^T V \gamma \neq 0\), unless \(U \perp V\),
- \(E (\hat{Y} - X\beta \mid X) = (P_U - I)V\gamma\), where \(P_U\) is proj. mat. for \(U\).
- (8) \(= \sigma^2 + \text{var}(\hat{m}_{LS}) + \text{bias}(\hat{m}_{LS}) = \sigma^2 + \text{rank}(U)\sigma^2 + \text{bias}\).
Correcting bias due to omitted important regressors

- Directly include $V$ in LM - infeasible (e.g., since it’s latent).
- Two established approaches:
  - 1925-28, P. Wright Instrumental variable(s):
    (i) Use observed instrument $Z$ to partial out endogeneity from $U$. Projecting $U$ on $Z \perp \varepsilon$ makes $\hat{\beta}$ unbiased.
    (ii) Pros: useful for inference when the omitted variable is not recoverable. Does not require normality.
    (iii) Cons: quality $\sim$ sample size, uncertainty for instrument.
  - 1976-79, J. Heckman Sample selection bias correction:
    (i) Use Probit model to recover the omitted variable from (censored) data and estimate the correct model.
    (ii) Pros: use full sample information, unbiased prediction.
    (iii) Cons: quality $\sim$ joint normality, inference misleading when $Y, V \sim U$. Induces heterogeneity in the model, which causes inefficiency.

Note: In 2000, Heckman shared Nobel Memorial Prize in Economic Sciences.
Correcting bias due to omitted important regressors

- Sample selection bias correction

Assume that the selected sample satisfies a LM

\[(0) \quad Y|(U, Y^* > 0) \sim (U\beta + f(Y^* > 0)\gamma, \sigma^2 I_n)\]

and additionally:

\[(I) \quad Y|U \sim (U\beta, \sigma^2_1 I_n), \text{ ok for population / full sample.}\]

\[(II) \quad Y^*|Z \sim (Z\beta_2, \sigma^2_2 I_n), \text{ ok for population / selected sample.}\]

Theoretically optimal choice of \(f(Y^* > 0)\) would be

\[
f(Y^* > 0)\gamma = E(\varepsilon(I)|Y^* > 0) = E(\varepsilon(I)|\varepsilon(II) > -Z\beta_2) \\ 
\overset{(III)}{=} E(\varepsilon(I)|\varepsilon(II) > -Z\beta_2) \\ 
\overset{N}{=} \frac{\sigma_{12}}{\sigma_2} \frac{\phi(-Z\beta_2/\sigma_2)}{1 - \Phi(-Z\beta_2/\sigma_2)} = \frac{\sigma_{12}}{\sigma_2} \lambda(Z, \beta_2, \sigma_2).
\]

Plugging into (0) we get

\[
Y|(U, Y^* > 0) \sim \left(U\beta + \lambda(Z, \beta_2, \sigma_2) \frac{\sigma_{12}}{\sigma_2}, \Omega_n\right),
\]

where \(\Omega_n\) is diagonal, but errors are heteroscedastic.

Note: The hazard function \(\lambda \downarrow 0\) for \(P(Y^* > 0) \uparrow 1\).
Correcting bias due to omitted important regressors
-Heckman’s estimation

(i) Missing $\lambda$, $Y^*$ $\rightarrow$ PROBIT on obs. $Z$ yields $\hat{\lambda}(\hat{\beta}_2, \hat{\sigma}_2)$. This estimation is consistent and based on full sample.

(ii) plugging $\hat{\lambda}$ into (0) and using OLS yields consistent but inefficient estimates $\hat{\beta}, \hat{\sigma}_{12}$ based on selected sample.

(iii) Estimates of $\sigma_1$ - see Heckman’s 1979 paper or GLS.

Notes from Heckman 1979:
(i) It is important for inference on $\hat{\beta}$ to obtain proper standard errors (usual proc. does not work).
(ii) Heckman proposed a testing procedure for presence of sample selection bias, i.e. $H_0 : \gamma = 0$.
(iii) R-package sampleSelection implements Heckman’s approach.
(iv) For a special case $Y = Y^*$ it becomes TOBIT model (J. Tobin, 1958).
(v) The estimates are not efficient, but can be used as starting values for ML estimation.
Reading for miscpecified linear models


iii Komarek Arnost, Linear regression, *Lecture notes - Charles University in Prague*, version 12/2018, link to homepage

iv Kulich Michal, Censored Data Analysis, *Lecture notes - Charles University in Prague*, version 1/2018, link to homepage

v Reschenhofer Erhard, Hilbert space geometry, *Lecture notes - University of Vienna*, link to homepage

Censored Heckman

-Maximum likelihood

Has this form (for one observation)

\[
L(\theta | y_o, y_s, x_S, x_O, T) = (9)
\]

\[
= \mathbb{P}(Y_o = NA)^{1-y_s} \mathbb{P}(Y_o = 0, Y_s = 1)^{y_s \{ y^*_o \leq T \}} f_{Y_o, Y_s}(\theta, y_o, y_s)^{y_s \{ y^*_o > T \}}
\]

\[
= (1 - \Phi(\beta_S^T x_S))^{1-y_s} \Phi_2 \left( \frac{T - \beta_O^T x_O}{\sigma}, \beta_S^T x_S, -\rho \right)^{y_s \{ y^*_o > T \}}
\]

\[
\cdots \cdot \left( \frac{1}{\sigma} \phi \left( \frac{y_o - \beta_O^T x_O}{\sigma} \right) \right) \Phi \left( \frac{\rho/\sigma (y_o - \beta_O^T x_O) + \beta_S^T x_S}{\sqrt{1 - \rho^2}} \right)^{y_s \{ y^*_o > T \}}
\]

where \( f \) is the joint density of \( Y_O, Y_S, \) and \( \phi, \Phi, \) and \( \Phi_2 \) are the pdf, cdf and joint cdf of the standard normal random variable(s with correlation coefficient \( \rho \)).

Data

- Descriptive summary - population 2015

profit/loss per employee
Data

- Descriptive summary - audits
Data

-Descriptive summary - population 2015

net assets per employee
Data

- Descriptive summary - audits

**net assets per employee**
Data

-Descriptive summary - population 2015

Average revenues per employee
Data
- Descriptive summary - audits

Average revenues per employee
Data

-Spacial distribution - audits
Data

- Spacial distribution - face-to-face audits
Data

- Descriptive summary - audits per ownership

\[ \log(\text{deficiency}+1) \]
Data

- Descriptive summary - audits per sector

$log(\text{deficiency}+1)$